Case Delicious Food

Lucas Victor Silva Pereira
lucasvsilvap@gmail.com

Cientista de Dados

1. Introdução

Este documento descreve o trabalho realizado na base de dados *DadosDeliciousFood.csv (DDF)*. Esta base de dados pertence a uma empresa *fictícia* do ramo alimentício, e para facilitar algumas explicações seu nome fictício é *Delicious Food (DF)*. A DF tem suas operações consolidadas na cidade do Rio de Janeiro e pretende expandir-se para a cidade de São Paulo. Dada a relevância do problema apresentado, este trabalho tem o objetivo de estudar a DDF para extrair informações que ajudem a DF a tomar a decisão mais assertiva possível.

A DDF é formada por uma combinação de informações públicas e privadas dos bairros das cidades do Rio de janeiro e São Paulo. As informações públicas, como tamanho da população e renda média, encontram-se no site do IBGE. Já informações como faturamento e potencial (de investimento ou lucro), são de propriedade da DF. A DDF é uma matriz bidimensional ($2D$) de tamanho 456×24, onde cada linha representa um bairro. Como a DF possui operações somente na cidade do Rio de Janeiro, as informações privadas, isto é, que advém do banco de dados da DF, só estão disponíveis para bairros do Rio de Janeiro. Note que todo o trabalho está concentrado exatamente em conseguir definir valores confiáveis para esses campos nos bairros da cidade de São Paulo. Em posse desses valores a DF terá uma segurança maior para tomar a decisão se deve ou não expandir. Portanto pode-se dividir a DDF em duas:

1. **DDF-RJ**: corresponde a 160 linhas da DDF e contém todos os bairros da cidade do Rio de Janeiro. É desta parte dos dados que serão extraídas as informações que ajudarão a classificar os bairros de São Paulo;
2. **DDF-SP**: corresponde ao restante das linha da DDF e contém todos os bairros da cidade de São Paulo presentes em DDF.

2. Tratamento dos Dados

A seção de tratamento dos dados foi dividida em duas: Seção 2.1 que apresenta a limpeza feita nos dados, e a Seção 2.2 que descreve de forma ampla algumas técnicas estatísticas utilizada para obter um maior conhecimento da base de dados.

2.1. Limpeza dos Dados

Para garantir que as informações obtidas terão boa qualidade, é preciso preparar a base de dados para então aplicar os algoritmos de aprendizagem. Inicialmente, é feita um limpeza

na base: retira-se os elementos repetidos, posteriormente é feito a verificação de dados faltantes e por fim, são retiradas as colunas que não agregam informação ao estudo.

Com relação aos dados faltantes, na DDF-RJ foram encontrados apenas seis campos vazios e na DDF-SP três, ambos na coluna renda média. Há algumas maneiras de resolver este problema, a título de exemplo, pode-se utilizar moda, mediana, a média calculada sobre os elementos não faltantes da coluna ou até mesmo métodos mais sofisticados para tentar prever esses valores, como uma regressão.

Contudo, dentre os 6 bairros do Rio de Janeiro que possuem renda média faltando, apenas um deles tem potencial médio, sendo todos os outros de baixo potencial para a DF. A decisão tomada para contornar o problema dos células vazias foi a exclusão das linhas. Entretanto, observe que não é o fato dos bairros possuírem baixo potencial para DF que levou a exclusão. Pensando nos algoritmos de aprendizado, eles precisam saber detectar quando um bairro possui um baixo potencial para a empresa. Eles aprendem a fazer isso analisando vários exemplos de bairros com baixo potencial. Portanto, o que levou a exclusão é o fato do número de bairros com baixo potencial ser maior do que bairros classificados com médio e alto potencial, ou seja, excluí-los evita a introdução de incertezas na DDF-RJ e contribui para tornar a base mais balanceada entre as três classes. O mesmo procedimento foi realizado na DDF-SP, excluir 3 de 296 linhas (bairros) não é uma perda considerável. Além dos bairros que possuíam campos vazios foi feito também uma checagem de dados inconsistentes, tal como, bairros com zero moradores, entre outras, mas nada foi encontrado.

Para finalizar a limpeza dos dados, cinco colunas foram retiradas, são elas:

- *Código*: este campo é um identificador único, conhecido como chave primária (*primary key*). É uma maneira de diferenciar uma tupla (linha) de todas as outras dentro da base de dados;
- *Nome*: nome do bairro;
- *Cidade*: cidade que o bairro pertence;
- *Estado*: estado que a cidade pertence;
- *População*: Quantidade total de moradores do bairro.

Os campos código e nome foram retirados pois não agregam informação útil a base de dados, ou seja, sabê-los não ajuda na definição do potencial de um determinado bairro para a DF. Já os campos cidade e estados possuem valores únicos dentro de cada uma das duas bases. Por fim, o mais complexo é a população. Como a base possui 8 campos que descrevem a população de forma discretizada, a informação que o campo população traz é redundante.

Após o término da limpeza das bases DDF-RJ ficou com 154 linhas e 19 colunas, já a DDF-SP possui 293 linhas e 17 colunas. A diferença no número de colunas entre DDF-RJ e DDF-SP acontece porque na base DDF-SJ não há valores de faturamento e potencial.

2.2. Analise das Variáveis

Após a limpeza dos dados é preciso verificar como se comportam as variáveis do problema. Nesta análise foram utilizados diversos artefatos estatísticos que estão explicados

em alto nível nos *scripts*[1] (códigos em python) disponibilizados juntos com este documento. Análise de estatística descritiva, correlação de Pearson, histogramas, gráficos de dispersão, boxplot, análise das componentes principais (PCA), entre outros.

O principal objetivo ao investigar as variáveis era obter como resultado, que a maioria deles descrevessem uma distribuição normal. Aceitando um certo número de *outliers* [2], este objetivo foi atingido.

Uma amostra será normalmente distribuída quando os dados dela podem ser descritos por uma função (curva) gaussiana. A Figura 1 apresenta o histograma da variável faturamento da base DDF-RJ. A linha azul no gráfico representa a função gaussiana. Quando a maioria das variáveis de uma base se comportam como uma distribuição normal, isso é um indicativo de que uma regressão linear consegue predizer dados relacionados à base com uma boa qualidade.

Histograma da Variável Faturamento

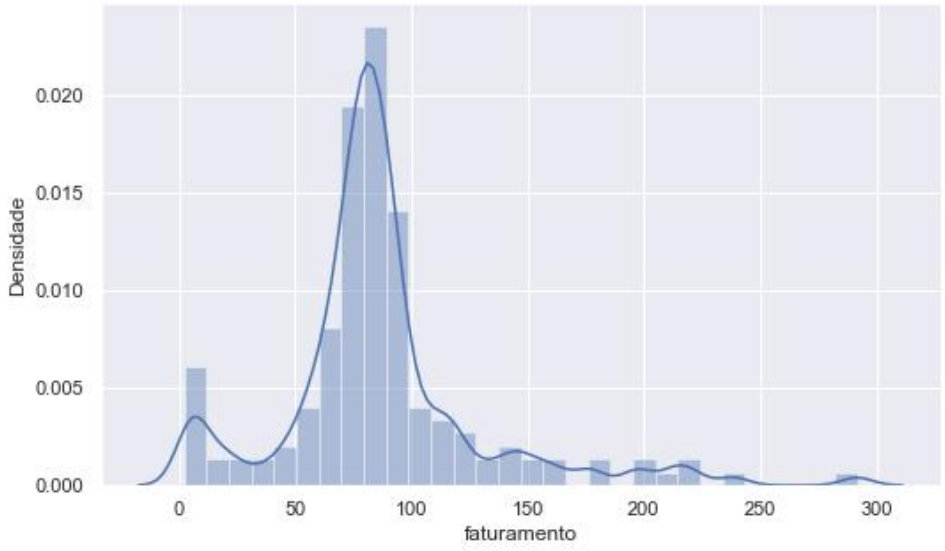

Figura 1. Observe que o gráfico é assimétrico à direita e possui *outliers*, por exemplo na primeira coluna. Contudo, se foi aceito um nível de erro, este histograma se comporta como uma distribuição normal.

3. Predizendo Valores de Faturamento para DDF-SP

Com o intuito de predizer os valores de faturamento aproximado que a DF terá caso resolva investir nos bairros de São Paulo, foi proposto um regressão linear como a apresentada abaixo:

$$y \approx \beta_3 x_3 + \beta_2 x_2 + \beta_1 x_1 + \beta_0$$

Os valores dos $\beta_{i's}$ são os parâmetros que o modelo deve aprender (parâmetros treináveis). Já os valores de $x_{i's}$ são quatro das principais variáveis que descrevem as características sobre o faturamento da DF em um determinado bairro. Os testes realizados apontaram que apenas quatro das vinte quatro variáveis que a DDF possui explicam mais de 90%

[1] Os códigos foram desenvolvido em *Jupyter Notebook*, assim foi possível conciliar um pouco de explicação entre as linhas de código.

[2] Dados amostrais que de destoam resto da amostra.

da variação do faturamento. A Figura 2 ilustra a qualidade da predição. A construção do gráfico implica em conhecer os valores reais do faturamento, por isso, o gráfico da Figura Figura 2 representa parte dos dados da base DDF-RJ [3].

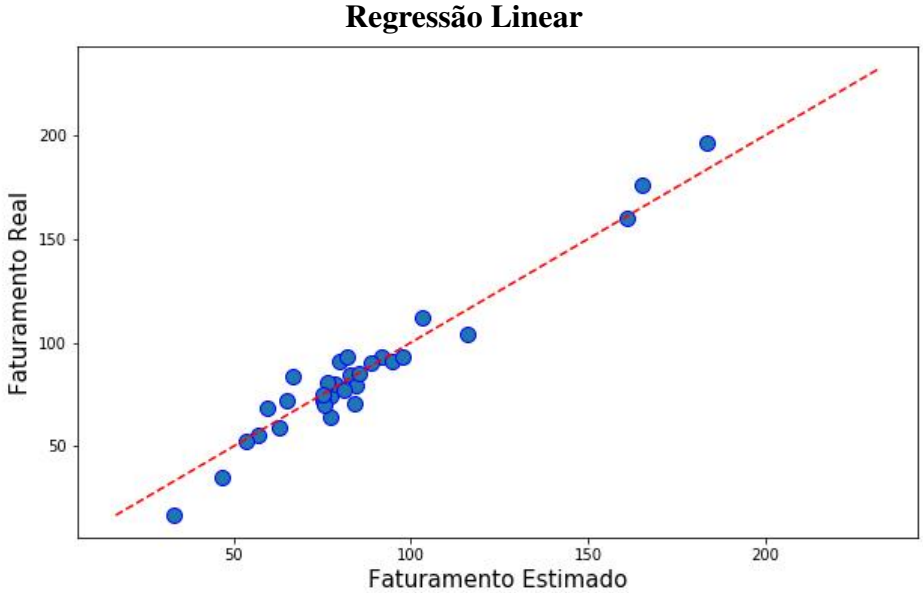

Figura 2. Cada ponto representa o valor predito (no eixo X), pelo valor real (no eixo Y). A reta vermelha corresponde à reta de regressão obtida com o treinamento do modelo.

4. Classificando o Potencial dos bairros da DDF-SP

A partir da função de regressão apresentada na seção anterior, os valores de faturamento da DDF-SP foram estimados e ala está preparada para ser classificada.

Novamente a base de dados DDF-RJ foi utilizada como referência na extração de informações com relação a variável potencial. O gráfico de dispersão da Figura 3 apresenta todos os bairros da base DDF-RJ. Perceba que utilizando apenas duas variáveis não é fácil separar os pontos pelas classes *baixo*, *médio* e *alto*, pois elas se misturam em determinas regiões do gráfico.

Como a informação da variável potencial importante, a estratégia utilizada para classificar a base DDF-SP, foi utilizar vários algoritmos de classificação, alguns deles são: *regressão logística*, *KNN*, *SVM*, *árvore de decisão*, *Random Forest* e outros.

A classificação é um problema mais fácil de ser resolvido do que a predição de valores contínuos (regressão), principalmente quando o número de classes é pequeno. Assumindo que o faturamento da DF pode ser negativo (quando ela tiver prejuízo em alguma filial, por exemplo), o modelo de regressão proposto pode atribuir qualquer valor à variável faturamento. Já o problema de classificação no caso da variável potencial, se restringe a decidir entre três possíveis classes. Diante disso, a maioria dos modelos possuem uma taxa de acurácia (acerto) superior a 80% nas amostras de teste. Os modelos baseados em árvore chegam a acurácias maiores que 90%.

[3]Especificamente, 20% da base DDF-RJ utilizada como conjunto de teste para treinar o modelo de regressão linear.

Figura 3. Os eixos são relacionados as duas variáveis mais correlacionadas com o faturamento. Os pontos em azul, vermelho e verde são classificados na base DDF-RJ como tendo respectivamente alto, médio e baixo potencial para DF.

Como a classificação dos bairros da base DDF-SP serão usados como base em decisões que envolvem muito dinheiro (como a expansão da DF), erros relacionados a falso positivo são muito piores do que os falsos negativos. Se um bairro qualquer da DDF-SP for classificado como tendo alto potencial, mas na verdade ele não pertence a está classe, pode levar a DF a fazer um investimento errado, ocasionado em uma perda financeira (econômica) significativa. Entretanto, um bairro que tem potencial alto mas não foi classificado como tal, pode fazer com que a DF deixa de deixe de faturar, mas ela não terá feito um investimento ruim, como no caso anteiro. Portanto, falso positivo e falso negativo devem ser evitados, mas o falso positivo tem um peso maior nesta comparação.

Afim de evitar os falsos positivos, a estratégia tomada foi utilizar mais de um classificador para dar o resultado final para à variável potencial, são eles: *KNN*, *SVM*, *Random Forest* e duas variações da *árvore de decisão*. A escolha dos modelos foi feita com base em dois tópicos: acurácia e diversificação. A acurácia é a definição mais utilizada para aferir o quão bom é um classificador. Já a diversificação foi levada em consideração porque o *ranking* dos classificadores propostos, dado pela acurácia, tem nas três primeiras posições algoritmos baseados em árvore de decisão. Apesar de serem diferentes, todos eles possuem o mesmo princípio, portanto, a chance deles cometerem os mesmos erros é alta. Por fim o critério de classificação utilizado foi:

- **Classe Alto**: quando todos os modelos classificam o bairro como sendo de alto potencial;
- **Classe Médio**: quanto nenhum dos modelos classifica o bairro como tendo baixo potencial, mas pelo menos um deles o classifica com potencial médio;
- **Classe Baixo**: quando pelo menos um dos classificadores identifica o bairro como tendo baixo potencial.

Como forma de visualizar a classificação de dois modelos: KNN e SVM, dois gráficos são apresentados na Figura 4. Observe pelas elipses pretas, em ambos os gráficos, que o classificador KNN não foi capaz de se moldar a ponto de conseguir classificar os

dois exemplos corretamente, algo que o SVM conseguiu. A seta preta apresenta um contorno entre as regiões da classe laranja (médio) e verde (alto) que pode ser um indicativo de *overfitting*[4], as cursvas do SVM são bem mais suaves. Entretanto, o SVM cria uma região da classe médio, no canto esquerdo superior que se estende até a parte de baixo, sem nem ter dados desta classe por perto. Note que o exemplo circulado em vermelho, provavelmente fez com que a região da classe verde se estendesse até lá. Isso mostra que ter uma base de dados balanceada e com exemplos diversos torna os classificadores mais robustos. Por fim, a estrategia de utilizar vários classificadores parece interessante, todos eles possuem pontos positivos e negativos.

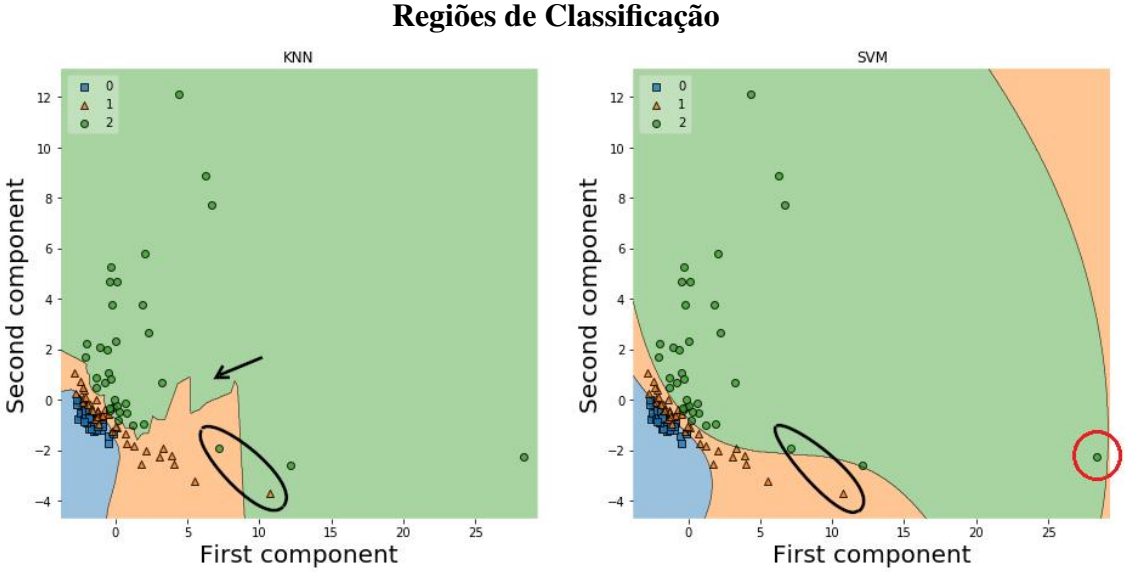

Figura 4. Os quadrados azuis, os triângulos laranja e as bolinhas verdes correspondem respectivamente as classes baixo, médio e Alto

Outro dado interessante que modelo Random Forest fornece é o gráfico de relevância da variáveis, apresentado na Figura 5. De acordo com o algoritmo, as variáveis ligadas a renda são mais importantes para descrever o potencial de um bairro do que aquelas ligadas a idade e tamanho da população. Observe também que as variáveis relacionadas a domicílios com renda mais alta e as variáveis ligada a população com 25 a 60 anos influenciam mais do que o restante das variáveis. Isso é um sinal de que a empresa esta conseguindo atingir seu publico alvo.

Dada a classificação da base DDF-SP é possível ordenar for faturamento, dentre os bairros que possuem potencial alto e então selecionar aquele que indica um maior faturamento como o primeiro bairro a receber uma sede da DF em São Paulo. A Tabela 1 apresenta os três primeiros bairros seguindo exatamente os passos descritos acima, realizado após a classificação da base DDF-SP. Além disso, é possível adicionar a está estratégia informações como maior incidência de publico alvo, por exemplo. Contudo, o próprio faturamento é um indicativo de onde esta o publico alvo. Outro fator relevante que trago para discussão com membros da DF é referente ao bairro *M'Boi Mirim*, ele possui a maior previsão de faturamento dentro todas os bairro nas dez simulações realizada. Entretanto, ela foi classificado como um bairro de potencial médio. Provavelmente, isso

[4]O overfitting ocorre quando os dados se ajustam demais aos dados de treinamento

se deve a um caso de falso negativo para o caso de potencial alto ou um erro referente a previsão feita pelo modelo de regressão linear.

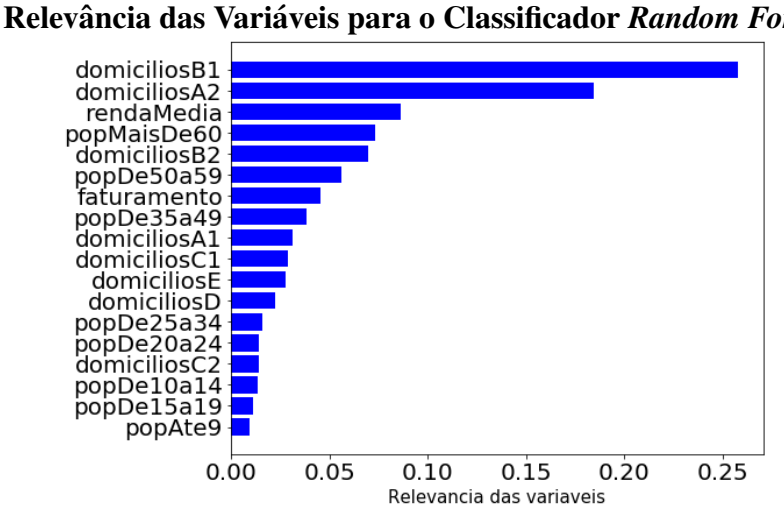

Figura 5. O gráfico apresenta a influência de cada uma das variáveis na estratégia do modelo Random Forest

Tabela 1. A tabela apresenta as três cidades que apareceram com maior frequência entre as três primeiras em um total de 10 simulações.

Nº	Bairros da Cidade de São Paulo
1	Pedreira
2	Tiquatira
3	Centro Empresarial

Todavia, dada que a classificação da variável potencial e a predição de valores para a variável faturamento foram feitas com base apenas em dados relacionados a idade e renda dos moradores de cada bairro, é interessante que algum profissional (possivelmente algum funcionário da DF) com conhecimento de domínio sobre o problema, valide os resultados obtidos neste trabalho.

5. Melhorias na Base de Dados

O objetivo principal do trabalho é classificar o potencial dos bairros da cidade de São Paulo. Um bairro terá alto potencial caso ele se mostre lucrativo para a empresa. Como o público alvo da empresa são pessoas de 25 a 50 anos com poder aquisitivo elevado, a maioria dos bairros classificados com alto potencial serão bairros nobres. Devido a isso é preciso levar em consideração custo, com por exemplo, aluguel do imóvel, deslocamento dos funcionários dentre outros que podem influenciar tanto no faturamento de forma direta como indireta. Existem algumas bases de dados que relacionam informações para predizer valores de aluguel na cidade de São Paulo.

Outros dados interessante são a distância entre os bairros e a própria localização de cada um deles. Um bairro com poucos moradores a princípio pode ser um desestímulo para a abertura de uma franquia da DF, contudo, se ele for caminho para vários pontos da

cidade, pode se tornar um atrativo, que pode inclusive esta relacionada a um aluguel mais barato. Além disso, é melhor a DF em um bairro de potencial alto que é perto de outros bairros de alto e médio potencial do que em um bairro afastado. Algumas das informações necessárias para realizar essas análises podem ser obtidas do google maps, por exemplo.

Por fim, também seria interessante utilizar os índices *IDH* e *Gini*. O índice de desenvolvimento humano está relacionado à qualidade de vida e o Gini à desigualdade social. Não sei afirmar se é fácil achar dados do índice Gini por bairros, mas o IDH é fácil encontrar para grandes cidades. É bem provável que tanto o IDH quanto o Gini estejam correlacionados com o faturamento, o primeiro de forma positiva e o segundo de forma negativa.